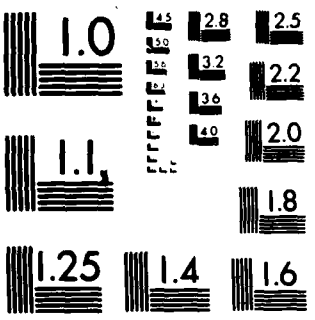
AD-A139 313 BOOTSTRAP INFERENCE WITH STRATIFIED SAMPLES(U) 1/1
WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER
J N RAO ET AL. JAN 84 MRC-TSR-2629 DAAG29-80-C-0041
UNCLASSIFIED F/G 12/1 NL



MRC Technical Summary Report #2629

BOOTSTRAP INFERENCE WITH STRATIFIED SAMPLES

J. N. K. Rao and C. F. J. Wu

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53705

January 1984

(Received December 21, 1983)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

BOOTSTRAP INFERENCE WITH STRATIFIED SAMPLES

J. N. K. Rao* and C. F. J. Wu

Technical Summary Report #2629
January 1984

ABSTRACT

The bootstrap method of inference is extended to stratified cluster samples when the parameter of interest, θ, is a nonlinear function $g(\bar{Y})$ of the population mean vector $\bar{Y}$. The bootstrap estimate of bias of $\hat{\theta} = g(\bar{y})$ and the estimate of variance of $\hat{\theta}$ are obtained, where $\bar{y}$ is a design-unbiased linear estimator of $\bar{Y}$. Bootstrap confidence intervals for θ are also given. Asymptotic justifications are provided.

AMS (MOS) Subject Classifications: 62D05, 62G15

Key Words: balanced or random half-samples, bootstrap, combined ratio, percentile method, stratified samples.

Work Unit Number 4 (Statistics and Probability)

* Carleton University, Canada

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

Most sample surveys involve stratification and multi-stage clustered sampling. A recent trend in survey data analysis is inference about nonlinear statistics from complex samples. Available methods include the linearization, jackknife and balanced half-samples. In the non-survey context, another method called the bootstrap has been shown to enjoy other desirable properties, the most important one being that it reflects the skewness inherent in the original point estimate. It is shown that a straightforward extension of the usual bootstrap provides incorrect variance estimates and misleading confidence intervals. A correct version is constructed by adjusting for a scaling problem before applying the nonlinear transformation. Several desirable theoretical properties of the proposed method are described. A detailed study in the special case of the combined ratio estimator is given.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the authors of this report.

A-1

BOOTSTRAP INFERENCE WITH STRATIFIED SAMPLES

J. N. K. Rao* and C. F. J. Wu

1. INTRODUCTION

Resampling methods, including the jackknife and the bootstrap, provide standard error estimates and confidence intervals for the parameters of interest. These methods are simple and straightforward but are computer-intensive, especially the bootstrap. Efron (1982) has given an excellent account of resampling methods in the case of an independent and identically distributed (i.i.d.) sample of fixed size n from an unknown distribution F, and the parameter of interest $\theta = \theta(F)$. The bootstrap confidence intervals for θ take account of the skewness in the estimator $\hat{\theta} = \theta(\hat{F})$, unlike the symmetric jackknife intervals based on the Student's t or the normal approximation. Moreover, limited empirical evidence (see Efron, 1982, p.18) has indicated that the bootstrap

* Carleton University, Canada

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

standard error estimates are likely to be more stable than those based on the jackknife and also less biased than those based on the customary delta (linearization) method. Holt and Scott (1983) applied the bootstrap to estimate variances of the regression estimators from data obtained from cluster sampling without stratification.

The main purpose of this article is to propose an extension of the bootstrap method to stratified samples, in the context of sample survey data; especially to data obtained from stratified cluster samples involving large numbers of strata, L, with relatively few primary sampling units (psu's) sampled within each stratum. For nonlinear statistics $\hat{\theta}$ that can be expressed as functions of estimated means of p (≥ 1) variables, Krewski and Rao (1981) established the asymptotic consistency of the variance estimators from the jackknife, the delta and the balanced repeated replication (BRR) methods as $L \to \infty$ within the context of a sequence of finite populations $\{\Pi_L\}$ with L strata in Π_L. Their result is valid for any multistage design in which the psu's are selected with replacement and in which independent subsamples are selected within those psu's sampled more than once. Rao and Wu (1983) obtained second order asymptotic expansions of these variance estimators under the above set up and made comparisons in terms of their biases.

The proposed bootstrap method for stratified samples is described in Section 2 and the properties of the resulting variance estimator are studied. The bootstrap estimate of bias of $\hat{\theta}$ is also obtained. Section 3 provides bootstrap confidence intervals for θ. The special case of a

ratio $\theta = \bar{Y}/\bar{X}$ is investigated in Section 4, where $\bar{Y}$ and $\bar{X}$ are the population means of variables y and x respectively. Finally, the results are extended to stratified simple random sampling without replacement in Section 5.

2. THE BOOTSTRAP METHOD

The parameter of interest θ is a nonlinear function of the population mean vector $\bar{Y} = (\bar{Y}_1,\ldots,\bar{Y}_p)^T$, say $\theta = g(\bar{Y})$. This form of θ includes ratios, regression and correlation coefficients. If n_h (≥ 2) psu's are selected with replacement with probabilities p_{hi} in stratum h, then Krewski and Rao (1981) have shown that the natural estimator $\hat{\theta} = g(\hat{\bar{Y}})$ can be expressed as $\hat{\theta} = g(\bar{y})$. Here $\hat{\bar{Y}}$ is a design-unbiased linear estimator of $\bar{Y} = \Sigma W_h \bar{Y}_h$ and $\bar{y} = \Sigma W_h \bar{y}_h$ where W_h and $\bar{Y}_h = (\bar{Y}_{h1},\ldots,\bar{Y}_{hp})^T$ are the h-th stratum weight ($\Sigma W_h = 1$) and population mean vector respectively and $\bar{y}_h$ is the mean of n_h i.i.d. random vectors $y_{hi} = (y_{hi1},\ldots,y_{hip})^T$ for each h with $E(y_{hi}) = \bar{Y}_h$. For $h \neq h'$, y_{hi} and $y_{h'j}$ are independent but not necessarily identically distributed.

2.1 The Naive Bootstrap

In the case of an i.i.d. sample $\{y_i\}_1^n$ with $E(y_i) = \bar{Y}$, the bootstrap method is as follows: (i) Draw a simple random sample $\{y_i^*\}_1^n$ with replacement from the observed values $y_1, y_2, \ldots, y_n$ and calculate $\hat{\theta}^* = g(\bar{y}^*)$ where $\bar{y}^* = \Sigma y_i^*/n$. (ii) Independently replicate step (i) a large number, B, of times and calculate the corresponding estimates $\hat{\theta}^{*1},\ldots,\hat{\theta}^{*B}$. (iii) The bootstrap variance estimator of $\hat{\theta} = g(\bar{y})$ is given by

$$v_b(a) = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}^{*b} - \hat{\theta}^*_a)^2 \qquad (2.1)$$

where $\hat{\theta}^*_a = \Sigma\hat{\theta}^{*b}/B$. The Monte-Carlo estimator $v_b(a)$ is an approximation to

$$v_b = \text{var}_*(\hat{\theta}^*) = E_*(\hat{\theta}^* - E_*\hat{\theta}^*)^2 \qquad (2.2)$$

where E_* denotes the expectation with respect to bootstrap sampling from a given sample $y_1,\ldots,y_n$. No closed-form expression for $\text{var}_*(\hat{\theta}^*)$ generally exists in the nonlinear case, but in the linear case with $p=1$, $\hat{\theta}^* = \bar{y}^*$ and v_b reduces to

$$\text{var}_*(\bar{y}^*) = \frac{n-1}{n^2} s^2 = \frac{n-1}{n} \text{var}(\bar{y}) \qquad (2.3)$$

where $(n-1)s^2 = \Sigma(y_i-\bar{y})^2$ and $\text{var}(\bar{y}) = s^2/n$ is the unbiased estimator of variance of $\bar{y}$. The modified variance estimator $[n/(n-1)]\text{var}_*(\hat{\theta}^*)$ exactly equals $\text{var}(\bar{y})$ in the linear case, but Efron (1982) found no advantage in this modification. In any case, $n/(n-1) \doteq 1$ in most applications and $\text{var}_*(\hat{\theta}^*)$ is a consistent estimator of the variance of $\hat{\theta}$, as $n \to \infty$ (Bickel and Freedman, 1981). The bootstrap histogram of $\hat{\theta}^{*1},\ldots,\hat{\theta}^{*B}$ may be used to find confidence intervals for θ that take account of the skewness in $\hat{\theta}$. This method (Efron, 1982) is called the percentile method.

Noting the i.i.d. property of the y_{hi}'s within each stratum, a straightforward extension of the previous bootstrap method to stratified samples is as follows: (i) Take a simple random sample $\{y^*_{hi}\}_{i=1}^{n_h}$ with replacement from the given sample $\{y_{hi}\}_{i=1}^{n_h}$ in stratum h, independently for each stratum. Calculate $\bar{y}^*_h = n_h^{-1}\Sigma_i y^*_{hi}$, $\bar{y}^* = \Sigma w_h \bar{y}^*_h$ and $\hat{\theta}^* = g(\bar{y}^*)$.

(ii) Independently replicate step (i) a large number, B, of times and calculate the corresponding estimates $\hat{\theta}^{*1},\ldots,\hat{\theta}^{*B}$. (iii) The bootstrap variance estimator of $\hat{\theta} = g(\bar{y})$ is given by

$$v_b(a) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{*b} - \hat{\theta}^{*}_a)^2 \tag{2.4}$$

where $\bar{y} = \Sigma W_h \bar{y}_h$ and $\hat{\theta}^{*}_a = \Sigma\hat{\theta}^{*b}/B$. The Monte-Carlo estimator $v_b(a)$ is an approximation to

$$v_b = \mathrm{var}_*(\hat{\theta}^*) = E_*(\hat{\theta}^* - E_*\hat{\theta}^*)^2 \tag{2.5}$$

where E_*, as before, denotes the expectation with respect to bootstrap sampling. In the linear case with p=1, $\hat{\theta}^* = \Sigma W_h \bar{y}^*_h = \bar{y}^*$ and v_b reduces to

$$\mathrm{var}_*(\bar{y}^*) = \Sigma \frac{W_h^2}{n_h}\left(\frac{n_h - 1}{n_h}\right)s_h^2 \tag{2.6}$$

where $(n_h-1)s_h^2 = \Sigma_i(y_{hi} - \bar{y}_h)^2$. Comparing (2.6) with the unbiased estimator of variance of $\bar{y}$, $\mathrm{var}(\bar{y}) = \Sigma W_h^2 s_h^2/n_h$, it immediately follows that $\mathrm{var}_*(\bar{y}^*)/\mathrm{var}(\bar{y})$ does not converge to 1 in probability, unless L is fixed and $n_h \to \infty$ for each h. Hence, $\mathrm{var}_*(\bar{y}^*)$ is not a consistent estimator of the variance of $\bar{y}$. It also follows that v_b is not a consistent estimator of the variance (or mean square error) of a general nonlinear statistic. There does not seem to be an obvious way to correct this scaling problem except when $n_h = k$ for all h in which case $k(k-1)^{-1}\mathrm{var}_*(\hat{\theta}^*)$ will be consistent. Bickel and Freedman (1983) also noticed the scaling problem, but they were mainly interested in bootstrap confidence intervals in the linear case (p=1). They have established the asymptotic N(0,1) property of the distribution of $t=(\bar{y}-\bar{Y})/[\mathrm{var}(\bar{y})]^{\frac{1}{2}}$

and of the conditional distribution of $(\bar{y}^* - y)/[\text{var}_*(\bar{y}^*)]^{\frac{1}{2}}$ in stratified simple random sampling with replacement, and also proved that $(\Sigma w_h^2 s_h^{*2}/n_h)/\text{var}_*(\bar{y}^*)$ converges to 1 in probability as $n = \Sigma n_h \to \infty$, where $(n_h-1)s_h^{*2} = \Sigma_i (y_{hi}^* - \bar{y}_h^*)^2$. Their result implies that one could use the bootstrap histogram of $\tilde{t}^{*1},\ldots,\tilde{t}^{*B}$ to find confidence intervals for $\bar{Y}$, where $\tilde{t}^{*b} = (\bar{y}^{*b} - \bar{y})/[\Sigma w_h^2 s_h^{*b2}/n_h]^{\frac{1}{2}}$ where s_h^{*b2} is the value of s_h^{*2} for the b-th bootstrap sample $(b = 1,\ldots,B)$. In the nonlinear case, there does not seem to be a simple way to construct $\tilde{t}^{*b}$-values similar to those of Bickel and Freedman since v_b has no closed form. Moreover, the straightforward extension of the bootstrap (hereafter called the naive bootstrap) does not permit the use of the percentile method based on the bootstrap histogram of $\hat{\theta}^{*1},\ldots,\hat{\theta}^{*B}$.

Although $\tilde{t}^{*b}$ is asymptotically $N(0,1)$ in the linear case, it is not likely to provide as good an approximation to the distribution of t as a statistic whose denominator and numerator are both adequate approximations to their counterparts in t. Such statistics will be proposed in Section 3.2. These are also applicable to the nonlinear case.

Recognizing the scaling problem in a different context, Efron (1982) suggested to draw a bootstrap sample of size $n_h - 1$ instead of n_h from stratum h $(h = 1,\ldots,L)$. In Section 2.2, we will instead propose a different method which includes his suggestion as a special case.

2.2. The Proposed Method

Our method is as follows: (i) Draw a simple random sample $\{y^*_{hi}\}_{i=1}^{m_h}$ of size m_h with replacement from $\{y_{hi}\}_{i=1}^{n_h}$. Calculate

$$\tilde{y}_{hi} = \bar{y}_h + m_h^{\frac{1}{2}}(n_h - 1)^{-\frac{1}{2}}(y^*_{hi} - \bar{y}_h)$$

$$\tilde{y}_h = m_h^{-1}\sum_{i=1}^{m_h}\tilde{y}_{hi} = \bar{y}_h + m_h^{\frac{1}{2}}(n_{h-1})^{-\frac{1}{2}}(\bar{y}^*_h - \bar{y}_h) \tag{2.7}$$

$$\tilde{y} = \Sigma W_h\tilde{y}_h \ , \quad \tilde{\theta} = g(\tilde{y}) \ .$$

(ii) Independently replicate step (i) a large number, B, of times and calculate the corresponding estimates $\tilde{\theta}^1,\ldots,\tilde{\theta}^B$. (iii) The bootstrap estimator $E_*(\tilde{\theta})$ of θ can be approximated by $\tilde{\theta}_a = \Sigma\tilde{\theta}^b/B$. The bootstrap variance estimator of $\hat{\theta}$ is given by

$$\tilde{\sigma}_b^2 = \tilde{v}_b = \mathrm{var}_*(\tilde{\theta}) = E_*(\tilde{\theta} - E_*\tilde{\theta})^2 \tag{2.8}$$

with its Monte-Carlo approximation

$$\tilde{\sigma}_b^2(a) = \tilde{v}_b(a) = \frac{1}{B-1}\sum_{b=1}^{B}(\tilde{\theta}^b - \tilde{\theta}_a)^2 \ . \tag{2.9}$$

One can replace $E_*\tilde{\theta}$ in (2.8) by $\hat{\theta}$.

2.3 Justification of the Method

In the linear case, $\theta = \bar{Y}$, $\tilde{v}_b$ reduces to the customary unbiased variance estimator $\mathrm{var}(\bar{y})$, noting that

$$\tilde{v}_b = E_*(\tilde{y} - \bar{y})^2 = \Sigma W_h^2\frac{m_h}{n_h - 1}\left\{\frac{1}{m_h}\left(\frac{n_h - 1}{n_h}\right)s_h^2\right\} = \Sigma\, W_h^2 s_h^2/n_h \ .$$

In the nonlinear case, we have shown in Appendix 1 that

$$\tilde{v}_b = v_L + O_p(n^{-2}) \ , \tag{2.10}$$

under the condition (A.1) given there.

where v_L is the linearization variance estimator:

$$v_L = \sum_{j,k=1}^{p} g_j(\bar{y}) g_k(\bar{y}) \{ \sum_{h=1}^{L} \frac{w_h^2}{n_h} s_{hjk} \} . \qquad (2.11)$$

where $g_j(t) = \partial g(t)/\partial t_k$ with $t = (t_1,\ldots,t_p)^T$ and $(n_h-1)s_{hjk} = \Sigma_i (y_{hij} - \bar{y}_{hj})(y_{hik} - \bar{y}_{hk})$ with $\bar{y}_{hj} = \Sigma_i y_{hij}/n_h$. In the linear case ($p = 1$), v_L reduces to $\text{var}(\bar{y})$. Under reasonable regularity conditions (see Appendix 1), v_L is a consistent estimator of variance of $\hat{\theta}$, $\text{Var}(\hat{\theta})$. Hence, it follows from (2.10) that $\tilde{v}_b$ is also consistent for $\text{Var}(\hat{\theta})$.

The asymptotic N(0,1) property of the conditional distribution of $(\tilde{\theta} - \hat{\theta})/\tilde{\sigma}_b$ can be established, assuming that $0 < \delta_1 \leq m_h/(n_h-1) \leq \delta_2 < \infty$ for all h , i.e. the bootstrap sample size m_h should be comparable to the original sample size n_h in each stratum. The proof is omitted since it follows along the lines of Bickel and Freedman (1983). This result provides an asymptotic justification for the percentile method based on the bootstrap histogram of $\tilde{\theta}^1,\ldots,\tilde{\theta}^B$ (Section 3.1 provides details of the percentile method).

2.4. Estimate of Bias of $\hat{\theta}$

Rao and Wu (1983) have shown that the bias of $\hat{\theta}$ is

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta = \frac{1}{2} \sum_{j,k=1}^{p} g_{jk}(\bar{Y}) \sum_{h=1}^{L} \frac{w_h^2}{n_h} S_{hjk} + \text{lower order terms} \qquad (2.12)$$

where $g_{jk}(\bar{Y})$ is the second derivative $\partial^2 g(t)/\partial t_j \partial t_k$ evaluated at

$t = \bar{Y}$ and $S_{hjk} = E(y_{hij} - \bar{Y}_{hj})(y_{hik} - \bar{Y}_{hk})$. Our bootstrap estimate of bias is

$$\tilde{B}(\hat{\theta}) = E_*(\tilde{\theta}) - \hat{\theta} \tag{2.13}$$

which is approximated by $\tilde{\theta}_a - \hat{\theta}$. From (A.2) and (A.3) of Appendix 1 we have

$$\tilde{B}(\hat{\theta}) = E_*(\tilde{\theta}) - \hat{\theta} = \frac{1}{2} \sum_{j,k=1}^{p} g_{jk}(\bar{y}) \sum_{h=1}^{L} \frac{W_h^2}{n_h} s_{hjk} + O_p(n^{-2}) \ . \tag{2.14}$$

Hence, $\tilde{B}(\hat{\theta})$ is a consistent estimator of $B(\hat{\theta})$. On the other hand, the bias estimate $\hat{B}(\hat{\theta}) = E_*(\hat{\theta}^*) - \hat{\theta}$, based on the naive bootstrap, is equal to

$$\hat{B}(\hat{\theta}) = \frac{1}{2} \sum_{j,k=1}^{p} g_{jk}(\bar{y}) \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left(\frac{n_h - 1}{n_h}\right) s_{hjk} + \text{lower order terms} \tag{2.15}$$

which is <u>not</u> a consistent estimator of $B(\hat{\theta})$. The proof of (2.15) is omitted since it follows along the lines of the proof of (2.14).

2.5. Choice of m_h

The choice $m_h = n_h$ is a natural one. The choice $m_h = n_h - 1$ gives $\tilde{y}_{hi} = y^*_{hi}$, and our method reduces to the naive bootstrap, except that, in step (i) of the latter method a simple random sample of size $n_h - 1$ is selected from $\{y_{hi}\}_{i=1}^{n_h}$ in stratum h. However, for n_h small it may not lead to stable variance estimators. For $n_h = 2$, $m_h = 1$, the method reduces to the well-known random half-sample replication and the resulting variance estimators are less stable than those obtained from BBR for the same number, B , of half-samples (McCarthy, 1969). For

the same number of pseudoreplications the bootstrap will in general perform less stably than the BRR, when the latter method is applicable. For $n_h \geq 4$, the choice $m_h = n_h - 1$ may be attractive since the variance estimators are likely to be more stable and since the naive bootstrap is being used. For small n_h, it may be worth considering a bootstrap sample size m_h slightly larger than n_h, say between n_h+1 and $2n_h$.

3. CONFIDENCE INTERVALS

We now consider different bootstrap methods for setting confidence intervals for θ.

3.1. Percentile Method

For ready reference, we now give a brief account of the percentile method based on the bootstrap histogram of $\tilde{\theta}^1, \ldots, \tilde{\theta}^B$. Define the cumulative bootstrap distribution function as

$$\hat{CDF}(t) = \#\{\tilde{\theta}^b \leq t \; ; \; b = 1, \ldots, B\}/B \; . \tag{3.1}$$

For $\alpha \leq 0.5$, define $\tilde{\theta}_{LOW}(\alpha) = \hat{CDF}^{-1}(\alpha)$ and $\tilde{\theta}_{UP}(\alpha) = \hat{CDF}^{-1}(1-\alpha)$. Then the interval

$$\{\tilde{\theta}_{LOW}(\alpha), \; \tilde{\theta}_{UP}(\alpha)\} \tag{3.2}$$

is an approximate $(1-2\alpha)$-level confidence interval for θ. It has the central $1-2\alpha$ portion of the bootstrap distribution (Efron (1982), p.78). One can also consider a bias-corrected percentile method, following Efron (1982, p.82). This method leads to

$$\{\hat{CDF}^{-1}(\Phi(2z_0 - z_\alpha)) \; , \; \hat{CDF}^{-1}(\Phi(2z_0 + z_\alpha))\} \tag{3.3}$$

as an approximate $(1-2\alpha)$-level confidence interval for θ, where Φ is the cumulative distribution function of a standard normal,

$z_0 = \Phi^{-1}(\widehat{CDF}(\hat{\theta}))$ and $z_\alpha = \Phi^{-1}(1-\alpha)$. The advantage of the interval (3.3) over (3.2) has been demonstrated by Efron (1982).

3.2. Bootstrap t-statistics

Instead of approximating the distribution of $\hat{\theta}$ by the bootstrap distribution of $\tilde{\theta}$, we can approximate the distribution of the t-statistic $t = (\hat{\theta}-\theta)/\tilde{\sigma}_b$ by its bootstrap counterpart $t^* = (\tilde{\theta} - \hat{\theta})/\tilde{\sigma}_b^*(a)$ where $\tilde{\sigma}_b^{*2}(a) = \tilde{v}_b^*(a)$ is the bootstrap variance estimator obtained from (2.9) by bootstrapping the particular bootstrap sample $\{\tilde{y}_{hi}\}$ i.e. by replacing y_{hi} by $\tilde{y}_{hi}$ in the proposed method. For the second phase bootstrapping one could choose values (m_h', B') different from (m_h, B). This double-bootstrap method thus leads to B values $t^{*1},\ldots,t^{*B}$ of t^*.

Utilizing the bootstrap histogram of $t^{*1},\ldots,t^{*B}$, we define $\widehat{CDF}_t(x) = \#\{t^{*b} \leq x\}/B$, $\tilde{t}_{LOW} = \widehat{CDF}_t^{-1}(\alpha)$, $\tilde{t}_{UP} = \widehat{CDF}^{-1}(1-\alpha)$, and construct an approximate $(1-2\alpha)$-level confidence interval for θ given by

$$\{\hat{\theta} - \tilde{t}_{UP}\tilde{\sigma}_b \ , \ \hat{\theta} - \tilde{t}_{LOW}\tilde{\sigma}_b\} \ . \tag{3.4}$$

The interval based on the t-statistic is likely to be better than the interval based on the percentile method (Babu and Singh, 1983).

We now provide an asymptotic justification for t^*. Noting that $\tilde{v}_b^*(a)$ is a Monte Carlo approximation to

$$\tilde{\sigma}_b^{*2} = \tilde{v}_b^* = E_{**}(\tilde{\theta}^* - E_{**}\tilde{\theta}^*)^2 \tag{3.5}$$

where $\tilde{\theta}^*$ is the value of $\tilde{\theta}$ obtained from bootstrapping the particular sample $\{\tilde{y}_{hi}\}$ and E_{**} is the second phase bootstrap expectation, we

can write $t^* = (\tilde{\theta} - \hat{\theta})/\tilde{\sigma}_b^*$. In the linear case $\theta = \bar{Y}$ it is easily seen that

$$E_* \tilde{\sigma}_b^{*2} = \tilde{\sigma}_b^2 \ . \tag{3.6}$$

In the nonlinear case, following Bickel and Freedman (1983), we can show that $\tilde{\sigma}_b^{*2}/\tilde{\sigma}_b^2$ converges to 1 in probability as $n \to \infty$. Hence, it follows that the conditional distribution of t^* is asymptotically N(0,1).

One could use a jackknife t-statistic $t_J = (\hat{\theta} - \theta)/\hat{\sigma}_J$ instead of t, where $\hat{\sigma}_J^2 = v_J$ is a jackknife variance estimator of $\hat{\theta}$ (see Krewski and Rao, 1981). The corresponding confidence interval is then given by

$$\{\hat{\theta} - \hat{t}_{UP}\hat{\sigma}_J, \ \hat{\theta} - \hat{t}_{LOW}\hat{\sigma}_J\} \tag{3.7}$$

where $\hat{t}_{LOW}$ and $\hat{t}_{UP}$ are the lower and upper α-points of the statistic $t_J^* = (\tilde{\theta} - \hat{\theta})/\hat{\sigma}_J^*$ obtained from the bootstrap histogram of $t_J^{*1}, \ldots, t_J^{*B}$, and $\hat{\sigma}_J^{*2}$ is obtained from $\hat{\sigma}_J^2$ by jackknifing the particular bootstrap sample $\{\tilde{y}_{hi}\}$. It can be shown that the confidence interval (3.7) is also asymptotically correct. A confidence interval of this type was considered by Efron (1981) in the case of an i.i.d. sample $\{y_i\}$.

It is also possible to replace $\hat{\sigma}_J^2$ by the BRR or the linearization variance estimator and obtain a confidence interval similar to (3.7).

4. COMBINED RATIO ESTIMATOR

The combined ratio estimator of the ratio $\theta = g(\bar{Y},\bar{X}) = \bar{Y}/\bar{X} = R$ (say) is given by $\hat{\theta} = g(\bar{y},\bar{x}) = \bar{y}/\bar{x} = r$ (say) where $\bar{y} = \Sigma W_h \bar{y}_h$ and $\bar{x} = \Sigma W_h \bar{x}_h$. The corresponding bootstrap estimator is $\tilde{\theta} = g(\tilde{y},\tilde{x}) = \tilde{y}/\tilde{x} = \tilde{r}$ (say), where

$$\tilde{\bar{y}} = \bar{y} + \Sigma W_h \sqrt{\frac{m_h}{n_h - 1}} \, (\bar{y}_h^* - \bar{y}_h) = \bar{y} + \Delta\bar{y}^*$$

and (4.1)

$$\tilde{\bar{x}} = \bar{x} + \Sigma W_h \sqrt{\frac{m_h}{n_h - 1}} \, (\bar{x}_h^* - \bar{x}_h) = \bar{x} + \Delta\bar{x}^* \, .$$

The bootstrap estimator of variance of r is given by

$$\begin{aligned} \tilde{v}_b &= E_*(\tilde{r} - r)^2 \\ &= v_L - \frac{2}{\bar{x}^3} \Sigma \frac{W_h^3}{n_h \sqrt{m_h (n_h - 1)}} s_{\hat{e}^2 xh} \\ &\quad + \frac{3}{\bar{x}^4}\left[(\Sigma \frac{W_h^2}{n_h} s^2_{\hat{e}h}) (\Sigma \frac{W_h^2}{n_h} s^2_{xh}) + 2(\Sigma \frac{W_h^2}{n_h} s_{\hat{e}xh})^2 \right] \\ &\quad + O_p(n^{-3}) \, , \end{aligned} \tag{4.2}$$

assuming that $\max_h W_h/n_h = O(n^{-1})$ and $0 < \delta_1 \le m_h/(n_h - 1) \le \delta_2 < \infty$ for all h (see Appendix 2). Here v_L is the linearization variance estimator

$$v_L = \frac{1}{\bar{x}^2} \Sigma \frac{W_h^2}{n_h} s^2_{\hat{e}h} \tag{4.3}$$

and

$$(n_h - 1) s_{\hat{e}^2 xh} = \sum_{i=1}^{n_h} \hat{e}^2_{hi} (x_{hi} - \bar{x}_h), \quad \hat{e}_{hi} = y_{hi} - \bar{y}_h - r(x_{hi} - \bar{x}_h) \, ,$$

and $s^2_{\hat{e}h}$, s^2_{xh}, $s_{\hat{e}xh}$ are respectively the sample variance of $\hat{e}_{hi}$, x_{hi} and the sample covariance of $\hat{e}_{hi}$ and x_{hi} in the h^{th} stratum. Since $s_{\hat{e}^2 xh} = 0$ for $n_h = 2$, [illegible]econd term of (4.2) is zero if $n_h = 2$ for all h. Since th[illegible]m of (4.2) is positive and of order $O_p(n^{-2})$, we have $\tilde{v}_b =$ [illegible]$^{-2})$ in general. On the other hand, the jackknife variance estimator satisfies $v_J = v_L + O_p(n^{-3})$ in

the special case of $n_h = 2$ for all h (Rao and Wu, 1983). The jackknife is too close to v_L in the latter case.

To obtain the bias of $\tilde{v}_b$, note that, when $e_{hi} = y_{hi} - \bar{y}_h - R(x_{hi} - \bar{x}_h)$ replaces its sample analaogue $\hat{e}_{hi}$ in (4.2), the only effect on (4.2) is that the error term is $O_p(n^{-2.5})$ instead of $O_p(n^{-3})$. By working on this modified formula of (4.2) and noting that $Es_{e^2xh} = (n_h-2)S_{e^2xh}/n_h$ where S_{e^2xh} is the population analogue of $s_{\hat{e}^2xh}$, we get the bias of $\tilde{v}_b$:

$$\text{Bias}(\tilde{v}_b) = \text{Bias}(v_L) - \frac{2}{\bar{x}^3}\Sigma \frac{W_h^3}{n_h^2}\frac{n_h-2}{[m_h(n_h-1)]^{\frac{1}{2}}} S_{e^2xh}$$

$$+ \frac{3}{\bar{x}^4}(\Sigma \frac{W_h^2}{n_h} s^2_{eh})(\Sigma \frac{W_h^2}{n_h} s^2_{xh}) + \frac{6}{\bar{x}^4}(\Sigma \frac{W_h^2}{n_h} S_{exh})^2 + O(n^{-3}) \quad (4.3)$$

$$= \text{Bias}(v_L) - 2a'' + 3b + 6c \ (say)$$

where S^2_{eh}, S^2_{xh} and S_{exh} are the population analogues of $s^2_{\hat{e}h}$, s^2_{xh} and $s_{\hat{e}xh}$ respectively. Using the result (Wu, 1982)

$$\text{Bias}(v_L) = -2a + b + O(n^{-3}) \quad (4.4)$$

where

$$a = \frac{1}{\bar{x}^3}\Sigma \frac{W_h^3}{n_h^2} S_{e^2xh},$$

we get from (4.3)

$$\text{Bias}(\tilde{v}_b) = -2a - 2a'' + 4b + 6c + O(n^{-3}). \quad (4.5)$$

In the special case of $n_h = 2$ for all h, $a'' = 0$ and

$$\text{Bias}(\tilde{v}_b) = \text{Bias}(v_{BRR-H}) = \text{Bias}(v_{BRR-S}) = -2a+4b+6c \qquad (4.6)$$

to $O(n^{-3})$, where v_{BRR-H} and v_{BRR-S} are the BRR variance estimators (see Rao and Wu, 1983). In the general case of $n_h \neq 2$ for at least one h, $\text{Bias}(\tilde{v}_b)$ depends on the bootstrap sample sizes $\{m_h\}$. In particular, if $m_h >> n_h$ we have $\text{Bias}(\tilde{v}_b) = -2a+4b+6c$ to $O(n^{-3})$. The choice $m_h = n_h - 1$ (Efron, 1982) leads to

$$\text{Bias}(\tilde{v}_b) = \text{Bias}(v_L) - \frac{2}{\bar{x}^3} \Sigma \frac{W_h^3(n_h-2)}{n_h^2(n_h-1)} S_{e^2xh} + 3b + 6c + O(n^{-3}) . \qquad (4.7)$$

5. STRATIFIED SIMPLE RANDOM SAMPLING WITHOUT REPLACEMENT

All the previous results apply to the case of stratified simple random sampling without replacement by making a slight change in the definition of $\tilde{y}_{hi}$:

$$\tilde{y}_{hi} = \bar{y}_h + m_h^{\frac{1}{2}}(n_h - 1)^{-\frac{1}{2}}(1 - f_h)^{\frac{1}{2}}(y^*_{hi} - \bar{y}_h) \qquad (5.1)$$

where $f_h = n_h/N_h$ is the sampling fraction in stratum h. It is interesting to observe that, even by choosing $m_h = n_h - 1$, $\tilde{y}_{hi} \neq y^*_{hi}$. Hence the naive bootstrap using y^*_{hi} will still have the problem of giving a wrong scale as discussed before. In the special case of $n_h = 2$ for all h, McCarthy (1969) used a finite population correction similar to (5.1) in the context of BRR.

Bickel and Freedman (1983) considered a different bootstrap sampling method in order to recover the finite population correction, $1 - f_h$, in the variance formula. This method essentially creates populations consisting of copies of each y_{hi}, $i = 1,\ldots,n_h$ and $h = 1,\ldots,L$

and then generates $\{y^*_{hi}\}_1^{n_h}$ as a simple random sample without replacement from the created population, independently in each stratum. This "blow-up" bootstrap was first proposed by Gross (1980) and also independently by Chao and Lo (1983). The variance estimator resulting from this method (by working directly with y^*_{hi}), however, remains inconsistent for estimating the true variance of $\hat{\theta}$. It is possible, however, to make the variance estimator consistent by reducing the bootstrap sample size to n_h-1, as in Section 2. In comparison with our method, the "blow-up" bootstrap is somewhat harder to implement, somewhat artificial and, if the stratum size, N_h, is not a multiple of n_h, requires an artificial randomization for choosing between two created populations.

ACKNOWLEDGEMENT

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. C.F.J. Wu is partially supported by NSF grant No. MCS-83-00140

APPENDIX

1. Proof that $\tilde{v}_b = v_L + O_p(n^{-2})$, assuming that

$$\max_h W_h/n_h = O(n^{-1}),\ 0 < \delta_1 \le m_h/(n_h-1) \le \delta_2 < \infty. \qquad (A.1)$$

The condition (A.1) allows L to be either bounded or unbounded.

Writing $\tilde{\bar{y}} - \bar{y} = \Delta\bar{y}^* = (\Delta\bar{y}^*_1, \ldots, \Delta\bar{y}^*_p)^T$, where $\Delta\bar{y}^*_j = \Sigma W_h \sqrt{\frac{m_h}{n_{h-1}}}\,(\bar{y}^*_{hj} - \bar{y}_{hj})$. we have

$$\mathrm{var}_*(\Delta\bar{y}^*_j) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} s^2_{hj} = O_p(n^{-1})$$

under the assumption (A.1) and the boundedness of $S_{hj}^2 = E(y_{hij} - \bar{Y}_{hj})^2$, i.e. $\max S_{hj}^2 < \infty$. Hence, $\Delta\bar{y}^* = O_p(n^{-\frac{1}{2}})$ and

$$\tilde{\theta} - \hat{\theta} = (\Delta\bar{y}^*)^T g'(\bar{y}) + \frac{1}{2}(\Delta\bar{y}^*)^T g''(\bar{y})(\Delta\bar{y}^*) + O_p(n^{-\frac{3}{2}}) \tag{A.2}$$

where $g'(\bar{y}) = (g_1(\bar{y}), \ldots, g_k(\bar{y}))^T$ and $g''(\bar{y})$ is the $p \times p$ matrix with elements $g_{jk}(\bar{y})$. Therefore,

$$\tilde{v}_b = E_*(\tilde{\theta} - \hat{\theta})^2 = \sum_{j,k=1}^{p} g_j(\bar{y}) g_k(\bar{y}) E_*(\Delta\bar{y}_j^*\Delta\bar{y}_k^*)$$

$$+ \sum_{j,k,\ell=1}^{p} g_j(\bar{y}) g_{k\ell}(\bar{y}) E_*(\Delta\bar{y}_j^*\Delta\bar{y}_k^*\Delta\bar{y}_\ell^*) + O_p(n^{-2}) .$$

Now noting that

$$E_*(\Delta\bar{y}_j^*\Delta\bar{y}_k^*) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} s_{hjk} \tag{A.3}$$

and

$$E_*(\Delta\bar{y}_j^*\Delta\bar{y}_k^*\Delta\bar{y}_\ell^*) = \sum_{h=1}^{L} W_h^3 \left(\frac{m_h}{n_h - 1}\right)^{\frac{3}{2}} E_*(\bar{y}_{hj}^* - \bar{y}_{hj})(\bar{y}_{hk}^* - \bar{y}_{hk})(\bar{y}_{h\ell}^* - \bar{y}_{h\ell})$$

$$= \sum_{h=1}^{L} \frac{W_h^3}{n_h^2} n_h [m_h(n_h - 1)]^{-\frac{1}{2}} s_{hjk\ell} = O_p(n^{-2})$$

under (A.1), we get the desired result. Here $(n_{h-1})s_{hjk\ell} = \sum_{i=1}^{n_h} (y_{hij} - \bar{y}_{hj}) \cdot (y_{hik} - \bar{y}_{hk})(y_{hi\ell} - \bar{y}_{h\ell})$.

2. Proof of (4.2). We follow the approach in Appendix 4 of Wu (1982) to derive (4.2). Under (A.1) we have $\Delta\bar{x}^*$, $\Delta\bar{y}^*$ and $\Delta\bar{e}^* = \Delta\bar{y}^* - r\Delta\bar{x}^*$ all of the order $O_p(n^{-\frac{1}{2}})$. Hence, noting that

$$r^* = r + \frac{\Delta\bar{e}^*}{\bar{x}}\left\{1 - \frac{\Delta\bar{x}^*}{\bar{x}} + \left(\frac{\Delta\bar{x}^*}{\bar{x}}\right)^2\right\} + O_p(n^{-2}) \ ,$$

we get

$$\tilde{v}_b = E_*(r^* - r)^2$$

$$= \bar{x}^{-2}\left[E_*(\Delta\bar{e}^*)^2 - 2E_*(\Delta\bar{e}^*)^2\frac{\Delta\bar{x}^*}{\bar{x}}\right. \tag{A.4}$$

$$\left. + 3E_*\frac{(\Delta\bar{e}^*\Delta\bar{x}^*)^2}{\bar{x}^2}\right] + O_p(n^{-3})$$

Now, writing $\Delta\bar{e}^* = \Sigma d_h\bar{e}_h^*$ where $d_h = W_h m_h^{-\frac{1}{2}}(n_h - 1)^{-\frac{1}{2}}$ and $m_h\bar{e}_h^* = \sum_{i=1}^{m_h} e_{hi}^* = \sum_{i=1}^{m_h}\{y_{hi}^* - \bar{y}_h - r(x_{hi}^* - \bar{x}_h)\}$, we get the following results:

$$E_*(\Delta\bar{e}^*)^2 = \sum_{h=1}^{L} d_h^2\frac{1}{m_h}\frac{1}{n_h}\sum_{i=1}^{n_h}\hat{e}_{hi}^2$$

$$= \sum_{h=1}^{L}\frac{W_h^2}{n_h}s_{\hat{e}h}^2 \tag{A.5}$$

$$E_*[(\Delta\bar{e}^*)^2\Delta\bar{x}^*] = \Sigma d_h^3 E_*[\bar{e}_h^{*2}(\bar{x}_h^* - \bar{x}_h)]$$

$$= \sum_{h=1}^{L}\frac{W_h^3}{n_h^2}\frac{n_h}{\sqrt{(n_h-1)m_h}}\cdot\frac{1}{n_h-1}\sum_{i=1}^{n_h}\hat{e}_{hi}^2(x_{hi} - \bar{x}_h)$$

$$= \sum_{h=1}^{L}\frac{W_h^3}{n_h^2}\frac{n_h}{\sqrt{(n_h-1)m_h}}s_{\hat{e}^2xh} \tag{A.6}$$

and

$$E_*[(\Delta\bar{e}^*)^2(\Delta\bar{x}^*)^2] = E_*[(\Sigma d_h^2\bar{e}_h^{*2})(\Sigma d_h^2(\bar{x}_h^* - \bar{x}_h)^2)] +$$

$$E_*[(\sum_{h\neq h'} d_h d_{h'}\bar{e}_h^*\bar{e}_{h'}^*)(\sum_{h\neq h'} d_h d_{h'}(\bar{x}_h^* - \bar{x}_h)(\bar{x}_{h'}^* - \bar{x}_{h'}))]$$

$$= [\Sigma d_h^2 E_*(\bar{e}_h^{*2})][\Sigma d_h^2 E_*(\bar{x}_h^* - \bar{x}_h)^2] +$$

$$2\sum_{h\neq h'} d_h^2 d_{h'}^2 E_*\bar{e}_h^*(\bar{x}_h^* - \bar{x}_h)E_*\bar{e}_{h'}^*(\bar{x}_{h'}^* - \bar{x}_{h'}) + O_p(n^{-3})$$

$$= (\Sigma \frac{d_h^2}{m_h}\frac{n_h - 1}{n_h} s_{\hat{e}h}^2)(\Sigma \frac{d_h^2}{m_h}\frac{n_h - 1}{n_h} s_{xh}^2) +$$

$$2[\sum_h d_h^2 E_*\bar{e}_h^*(\bar{x}_h^* - \bar{x}_h)]^2 + O_p(n^{-3})$$

$$= (\Sigma \frac{w_h^2}{n_h} s_{\hat{e}h}^2)(\Sigma \frac{w_h^2}{n_h} s_{xh}^2) + 2(\Sigma \frac{w_h^2}{n_h} s_{\hat{e}xh})^2 + O_p(n^{-3}). \quad (A.7)$$

Substituting (A.5)-(A.7) in (A.4) we get the desired result.

REFERENCES

Babu, G.J. and Singh, K. (1983) Inference on means using the bootstrap. Ann. Statist., 11, 999-1003.

Bickel, P.J. and Freedman, D.A. (1981) Some asymptotic theory for the bootstrap. Ann. Statist., 9, 1196-1217.

Bickel, P.J. and Freedman, D.A. (1983) Asymptotic normality and the bootstrap in stratified sampling. Preprint.

Chao, M.T. and Lo, S.H. (1983) A bootstrap method for finite population. Preprint.

Efron, B. (1981) Nonparametric standard errors and confidence intervals. Canadian J. Statist., 9, 139-172.

Efron, B. (1982) The Jackknife, the Bootstrap and Other Resampling Plans. SIAM, Philadelphia.

Gross, S. (1980) Median estimation in sample surveys. Proc. Amer. Statist. Assoc., Section on Survey Research Methods, 181-184.

Holt, D. and Scott, A.J. (1983) Variances of regression estimators with survey data using the bootstrap or jackknife. 44th Session of the International Statistical Institute: Contributed Papers, 337-341.

Krewski, D. and Rao, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. Ann. Statist., 9, 1010-1019.

McCarthy, P.J. (1969) Pseudoreplication: half samples. Rev. Int. Statist. Inst., 37, 239-264.

Rao, J.N.K. and Wu, C.F.J. (1983) Inference from stratified samples: second order analysis of three methods for nonlinear statistics. Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University, Ottawa, No. 7.

Wu, C.F.J. (1982) Variance estimation for the combined ratio and combined regression estimators. Preprint.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 2629	2. GOVT ACCESSION NO. AD-A139 313	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) BOOTSTRAP INFERENCE WITH STRATIFIED SAMPLES		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) J. N. K. Rao and C. F. J. Wu		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE January 1984 13. NUMBER OF PAGES 20
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

balanced or random half-samples, bootstrap, combined ratio, percentile method, stratified samples

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The bootstrap method of inference is extended to stratified cluster samples when the parameter of interest, θ, is a nonlinear function, $g(\bar{Y})$ of the population mean vector $\bar{Y}$. The bootstrap estimate of bias of $\theta = g(\bar{y})$ and the estimate of variance of θ are obtained, where $\bar{y}$ is a design-unbiased linear estimator of $\bar{Y}$. Bootstrap confidence intervals for θ are also given. Asymptotic justifications are provided.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)